A new parametrizable model of molecular electronic structure

Dimitri N. Laikov[*]

Chemistry Department, Moscow State University, 119992 Moscow, Russia

(Dated: August 30, 2011)

Abstract

A new electronic structure model is developed in which the ground state energy of a molecular system is given by a Hartree-Fock-like expression with parametrized one- and two-electron integrals over an extended (minimal + polarization) set of orthogonalized atom-centered basis functions, the variational equations being solved formally within the minimal basis but the effect of polarization functions being included in the spirit of second-order perturbation theory. It is designed to yield good dipole polarizabilities and improved intermolecular potentials with dispersion terms. The molecular integrals include up to three-center one-electron and two-center two-electron terms, all in simple analytical forms. A method to extract the effective one-electron Hamiltonian of nonlocal-exchange Kohn-Sham theory from the coupled-cluster one-electron density matrix is designed and used to get its matrix representation in a molecule-intrinsic minimal basis as an input to the paramtrization procedure – making a direct link to the correlated wavefunction theory. The model has been trained for 15 elements (H, Li–F, Na–Cl, 720 parameters) on a set of 5581 molecules (including ions, transition states, and weakly-bound complexes) whose first- and second-order properties were computed by the coupled-cluster theory as a reference, and a good agreement is seen. The model looks promising for the study of large molecular systems, it is believed to be an important step forward from the traditional semiempirical models towards higher accuracy at nearly as low a computational cost.

---

[*]E-Mail: laikov@rad.chem.msu.ru; Homepage: `http://rad.chem.msu.ru/~laikov/`

I. INTRODUCTION

Atomistic computer simulations of complex chemical systems and materials at quantitative level are a great challenge for modern science: not only a high accuracy of computed potential energy surfaces is needed for systems with many atoms, but also a higher speed of computation for a thorough sampling of the configurational space. Molecular mechanical force fields pioneered [1] 44 years ago are still almost the only practical method of calculation in many fields thanks to their very high speed and despite their well-known limitations. With their fixed bonding topology, they are designed first of all for conformational studies where nonbonding interactions play the main role, their treatment of electrostatics may be as sophisticated as to account for polarization effects [2]. Extensions to treat one (or few) chemical reaction steps [3] need careful parameter adjustment for each active center studied. General reactive force fields [4, 5] with geometry-dependent bond orders and atomic charges seem to be a logical next step and already remind of models within the Hohenberg-Kohn [6] density functional theory (DFT). Our own experience with this kind of models lead us to believe that instead of designing more and more complicated charge and bond order functionals it should be much easier to incorporate at least a single matrix diagonalization of an effective one-electron Hamiltonian [7] into the model (likely in a linear-scaling fashion [8, 9]), just like the mainstream DFT took the Kohn-Sham path [10], and thus we escape the force field and enter the realm of molecular quantum mechanics.

At the other extreme, rigorous wavefunction methods give useful results starting with the second-order many-body perturbation theory (MP2) [11], but much better with the coupled-cluster theory [12] with single and double substitutions (CCSD) [13] or – as the golden standard – with the further perturbative account of triple substitutions (CCSD(T)) [14]. Their fifth, sixth, and seventh power scaling of computational cost with the system size and huge prefactors, as well as the slow convergence of the computed properties with the basis set size, make them hopelessly slow for large molecules, but they are indispensable for getting small-molecule reference data for training and testing all kinds of models.

Kohn-Sham DFT on its way from local [10] to generalized-gradient approximations [15, 16], with further inclusion of a fraction [17] or, much better, the full long-range part [18] of the nonlocal exchange and the dispersion tail [19, 20], has grown into a rather accurate electronic structure model with favorable system-size scaling properties. In a standard implementation,

2

localized atom-centered basis functions are used to solve the self-consistent field (SCF) equations (plane wave techniques [21] are limited to the less accurate pure density functionals and cannot work as fast with the nonlocal exchange), the analytical evaluation of two-electron Coulomb integrals and the numerical integration of exchange-correlation terms are the bottleneck for smaller system sizes but can be made linear scaling starting from around 1000 atoms, in that limit we estimate the computation of DFT energy functional to be about 10^6 to 10^8 times slower than the most sophisticated polarizable force field energy and gradient evaluation. Density-fitting techniques [22, 23] can speed up the calculation of these terms by up to 100 times, but only for the pure DFT. Pseudospectral methods [24, 25] can show up to 100 times speed-up also for the nonlocal-exchange DFT. Even if the integrals were for free, there would be another serious bottleneck in the linear algebra of SCF equations. A smallest meaningful atomic basis set with 5 functions for H and 14 or 18 for Li–Ne would yield little to no sparsity of the density matrix for typical three-dimensional molecules with up to 3000 atoms [26] and around 30000 basis functions, so about 10^{13} floating-point multiplications and additions would be needed to do one matrix-multiply! In this regime, one SCF energy calculation can hardly take much less than one day even on a modern high-performance parallel computer.

The only way for the electronic structure theory to compete with the force field methods lies in the use of a finite-dimensional model Hamiltonian defined by its matrix elements in a minimal atomic basis representation. This is an alternative strategy with the Kohn-Sham DFT – instead of defining the universal density functional in some limited (approximate, parametrized) form and then computing the arising integrals rigorously at each molecular geometry, these molecular integrals themselves can be directly modeled as the functions of atomic coordinates. One may note that the Coulomb potential of the nuclei in molecules is a very special class of external potentials for a system of electrons, and the widely-used density functionals are not as universal as thought, being often fitted [27] to molecular data. The minimal basis representation is not a limitation – not a fixed free-atom but a molecule-adapted set of (deformed) atomic functions is implied. In our earlier work [28] we have shown how to extract such a basis set from an accurate (or exact) solution of Kohn-Sham equations at a given molecular geometry in a unique and molecule-intrinsic way. The effective local potential of Kohn-Sham theory can in turn be extracted from the correlated wavefunction theory by a stable numerical procedure [29], in Section II D of this work we

derive its analog for the case of nonlocal exchange. (After we had done ours, we were made aware [30] of a parallel work on the effective valence shell Hamiltonians [31–34] but for a fully *correlated* – and not SCF – treatment of the ground and excited states). Now that the effective minimal-basis matrix elements of the model Hamiltonian have the first-principles foundation and can be computed numerically, they can be used to guide the work on their approximations.

Into this category fall the semiempirical electronic structure models based on the neglect of diatomic differential overlap (NDDO) formalism that have had nearly half a century of conservative evolution. Conceived as the simplest approximations [35] for valence-only minimal-basis SCF calculations with most multicenter integrals set to zero, they became a breakthrough in computational chemistry [36] when a systematic parametrization [37–42] to fit experimental molecular data turned them into predictive phenomenological models. As the growing computer power allowed the DFT, MP2, and other rigorous methods to be applied to chemically-interesting systems and the standards of accuracy tightened, the semiempirical models began to lose the game. Some limited orthogonalization corrections were studied [43, 44] but lead only to a limited improvement. Poor intermolecular potentials, especially for hydrogen-bonding, were tried to be cured [45, 46] by simple diatomic dispersion corrections [47] and further by triatomic corrections [48, 49] in the spirit of molecular mechanics – by adding the terms that depend on the atomic coordinates only and not on the electronic state. A proper treatment of polarizability within the minimal-basis formalism is not straightforward, one solution [50] is to add polarization functions into the basis but this slows down the calculations. A more attractive way was found [51, 52] in which a polarization term borrowed from molecular mechanics is added on top of the SCF equations. With all these recent developments, the desired improvement in accuracy of the semiempirical NDDO models for many applications is still not achieved.

Here we report our new parametrizable electronic structure model that evolves along a different line. We consider an extended atomic valence basis set that has radial and angular polarization functions added to the minimal set on each atom, and then we design a new two-layer SCF method (Section II A) in which the variational equations are defined within the minimal subspace and the contribution from the polarization subspace is included in the spirit of second-order perturbation theory. It should be noted that our SCF method is fairly different in many ways from other known dual-basis techniques [53–55]. The parametriz-

able molecular integrals of our model (Section II B) have more general functional forms and include new terms not seen in the traditional semiempirical NDDO models. We add the long-range dispersion corrections into the two-electron integrals – a more meaningful treatment of the two-electron correlation effect that depends on the electronic state. Our model also treats the molecular polarizability in a natural and self-interaction-free way. The parametrization procedure (Section II C) we use to train our model has a number of new terms added into the optimization process: the effective one-electron Hamiltonian matrices extracted (Section II D and [28]) from the correlated wavefunction theory, the molecular dispersion coefficients (Section II E) and the molecular electrostatic potentials on the surfaces (Section II F), as well as the second-order response properties (force constant matrices, atomic polar tensors, and dipole polarizabilities) – all these values coming from the high-level (coupled-cluster) calculations are required to be reproduced (in the least-squares sense) by the parametrized SCF model.

All this methodology defines our model up to the values of parameters that have to be optimized on a set of reference molecular data. Here we report (Section III) a preliminary parametrization for 15 elements H, Li–F, Na–Cl on a very diverse set of molecular structures to assess the worst-case performance of the model – it stands the test, seems to be already usable and useful for some applications, and awaits a future extension to heavier elements.

II. THEORY

A. Energy expression

In our model, the total energy of a molecular system with K nuclei at positions $\{\mathbf{r}_k\}$, $k = 1, \ldots, K$, in an external field $v(r)$ can be written as a sum of three terms

$$E = E_0 + E_1 + E_2, \tag{1}$$

the one independent of the valence electronic structure

$$E_0 = \sum_k e_k + \sum_{k<k'} q_k q_{k'} |\mathbf{r}_k - \mathbf{r}_{k'}|^{-1} + \sum_k q_k v(\mathbf{r}_k) \tag{2}$$

being the sum of core electron energies e_k, the repulsion of atomic cores with charges $\{q_k\}$ and their interaction with the external field; the Hartree-Fock energy within the minimal

atomic basis

$$E_1 = \tfrac{1}{2} \sum_{\mu\nu\sigma} \left(H_{\mu\nu} + F_{\mu\nu\sigma} \right) D_{\mu\nu\sigma}, \tag{3}$$

$$F_{\mu\nu\sigma} = H_{\mu\nu} + \sum_{\mu'\nu'\sigma'} R_{\mu\nu\sigma\mu'\nu'\sigma'} D_{\mu'\nu'\sigma'}, \tag{4}$$

the indices $\mu, \nu = 1, \ldots, M$ running over the minimal set, the spin label $\sigma = \pm\tfrac{1}{2}$; and the second-oder term

$$E_2 = \sum_{\mu\nu\sigma} F^{(2)}_{\mu\nu\sigma} D_{\mu\nu\sigma}, \tag{5}$$

$$F^{(2)}_{\mu\nu\sigma} = \sum_{\alpha} \left(F_{\alpha\mu\sigma} X_{\alpha\nu\sigma} \bar{s}_{\nu\sigma} + F_{\alpha\nu\sigma} X_{\alpha\mu\sigma} \bar{s}_{\mu\sigma} \right), \tag{6}$$

$$X_{\alpha\mu\sigma} = \frac{F_{\alpha\mu\sigma}}{\varepsilon_\mu - \varepsilon_\alpha}, \tag{7}$$

$$F_{\alpha\nu\sigma} = H_{\alpha\nu} + \sum_{\mu'\nu'\sigma'} R_{\alpha\nu\sigma\mu'\nu'\sigma'} D_{\mu'\nu'\sigma'}, \tag{8}$$

that accounts for the effect of polarization functions (labeled by $\alpha = M + 1, \ldots, N$) in a perturbative way. The one-electron density matrices

$$D_{\mu\nu\sigma} = \sum_{i=1}^{N_\sigma} C_{\mu i\sigma} C_{\nu i\sigma} \tag{9}$$

with N_σ electrons for each spin are formed from the orthogonal coefficient matrices

$$\sum_{\mu} C_{\mu i\sigma} C_{\mu j\sigma} = \delta_{ij}. \tag{10}$$

The orthogonalized atomic valence basis functions

$$\phi_\mu = \phi_{m_\mu l_\mu n_\mu k_\mu} \tag{11}$$

can also be labeled by their principal n, angular l, and azimuthal m quantum numbers and atomic centers k, for each l no more than one n value is used in either minimal or polarization set. The one-electron integral matrix is a sum

$$H_{\mu\nu} = T_{\mu\nu} + \sum_{k'} V_{\mu\nu k'} + V^{(v)}_{\mu\nu} \tag{12}$$

of kinetic energy, effective core potential, and external field terms. The spin-dependent two-electron integrals

$$R_{\mu\nu\sigma\mu'\nu'\sigma'} = R_{\mu\nu\mu'\nu'} - \delta_{\sigma\sigma'} R_{\mu'\nu\mu\nu'} \tag{13}$$

6

are built from the spinless Coulomb repulsion integrals by the proper antisymmetrization. The values $\varepsilon_\mu < 0$ and $\varepsilon_\alpha > 0$ in the denominator of Eq. (7) are taken as atomic constants and play the role of diagonal energies of the zeroth-order problem of perturbation theory, their signs imply that $E_2 \leq 0$. Had we had $\bar{s}_{\mu\sigma} \equiv 1$ in Eq. (6) as we did in an earlier model, the total energy E would be a cubic function of the density matrix, whereas the Hartree-Fock energy E_1 alone is quadratic – the perturbative inclusion of higher-order basis set effects is leading to the appearance of effective three-electron integrals. Moreover, E_2 would be quadratic in the external field – as needed for the proper account of polarizability within a formally minimal-basis treatment. Our experience has shown, however, that the simplest choice of $\bar{s}_{\mu\sigma} \equiv 1$ often leads to a collapse of E_2 when atoms in a molecule get crowded – we have overcome this problem using the damping factors

$$s_{\mu\sigma} = \left(1 + \sum_\alpha X_{\alpha\mu\sigma}^2\right)^{-1} \tag{14}$$

that need to be spherically averaged

$$\bar{s}_{\mu\sigma} = \frac{1}{1 + 2l_\mu} \sum_\nu s_{\nu\sigma} \delta_{l_\mu l_\nu} \delta_{k_\mu k_\nu} \tag{15}$$

before the insertion into Eq. (6).

It is noteworthy that our model has the integrals of only two types: either with all functions from the minimal set (all-minimal) as in Eq. (4) or with all but one from the minimal and one from the polarization set (minimal-polarization one-electron and 3-minimal-1-polarization two-electron integrals) as in Eq. (8).

The energy expression (1) is minimized under the orthogonality constraints (10) to get the ground-state solution of the electronic structure problem, the stationary conditions can be given in the form of self-consistent-field (SCF) one-electron equations

$$\sum_\nu \left(\mathcal{F}_{\mu\nu\sigma} - \delta_{\mu\nu}\epsilon_{i\sigma}\right) C_{\nu i\sigma} = 0 \tag{16}$$

with the effective Hamiltonian matrix derived as

$$\mathcal{F}_{\mu\nu\sigma} = \partial E / \partial D_{\mu\nu\sigma}, \tag{17}$$

the explicit expression for the latter is lengthy but straightforward. To compute a one-electron property V such as dipole moment or molecular electrostatic potential in a consistent

way, the energy derivative formalism should be used leading to

$$V = \sum_{\mu\nu\sigma} V_{\mu\nu} D_{\mu\nu\sigma} + \sum_{\alpha\nu} V_{\alpha\nu} D_{\alpha\nu} \tag{18}$$

with the minimal-polarization block of the effective density matrix given by the derivative

$$D_{\alpha\nu} = \partial E / \partial H_{\alpha\nu}, \tag{19}$$

its explicit form is again lengthy but straightforward.

We have written a computer code that solves these SCF equations to get the energy and its first and second derivatives with respect to the atomic coordinates and the applied uniform electric field, the derivatives being computed in a fully analytic way.

B. Molecular integrals

The one- and two-electron integrals in the energy expression (1) refer to the orthogonalized atom-centered basis functions (11) and as such are nontrivial functions of (in general) all atomic coordinates of the molecular system. They could have been computed from the first principles using a three-level extension of our zero-bond-dipole orthogonalization scheme [28] applied to a fixed all-electron atomic basis – the core set is orthogonalized first, the minimal valence set is orthogonalized to the core and then within itself, and the polarization set is orthogonalized to the core and valence and at last within itself. Such first-principles integrals, however, can give, at best, only a rough approximation to the Hartree-Fock molecular energy within a polarized basis, but we are aiming at a *parametrizable* model that can yield accurate molecular properties with the electron correlation included in the spirit of Kohn-Sham density-functional theory. Thus our model uses parametrized explicit analytical formulas for the molecular integrals giving values that differ somewhat from the first-principles ones.

The two-electron integrals are split into a sum

$$R_{\mu\nu\mu'\nu'} = R^{(0)}_{\mu\nu\mu'\nu'} + R^{(6)}_{\mu\nu\mu'\nu'} \tag{20}$$

of "electrostatics" and "dispersion" parts, the former quickly (exponentially) reaching the asymptotics of Coulomb interaction of the two ($\mu\nu$ and $\mu'\nu'$) charge distributions and the latter having a characteristic r^{-6} asymptotic tail. For the integrals with one polarization function

$$R_{\alpha\nu\mu'\nu'} = R^{(0)}_{\alpha\nu\mu'\nu'} \tag{21}$$

only the electrostatic part is used. The electron correlation can already be modeled by decreasing the magnitude of the two-electron integrals.

The one-electron integrals (12) have long-range Coulomb terms $V_{\mu\nu k'}$ and it is much easier to work with their short-range analogs

$$\bar{F}_{\mu\nu} = H_{\mu\nu} + \sum_{\mu'} \left(R^{(0)}_{\mu\nu\mu'\mu'} - \tfrac{1}{2} R^{(0)}_{\mu\mu'\mu'\nu} \right) p_{\mu'}, \tag{22}$$

and in the same way for $\bar{F}_{\alpha\nu}$, where a diagonal *promolecule* density matrix with constant spherically-symmetric atomic populations

$$p_\mu = p_{l_\mu k_\mu} \tag{23}$$

is used to add the promolecule Coulomb and exchange terms, so that they fully neutralize the core charges

$$\sum_l p_{lk} = q_k. \tag{24}$$

It is trivial to rewrite the energy expression in terms of $\bar{F}$ and R integrals, working with the short-range integrals not only simplifies the parametrization but is also of great advantage for large-scale calculations. Moreover, some two-electron terms can be accounted for (on the average) only within the $\bar{F}$ integrals and neglected in the R integrals, as we do in the following. The (promolecule) one-electron integrals $\bar{F}_{\mu\nu}$ are of two kinds: one-center if $k_\mu = k_\nu$, and two-center if $k_\mu \neq k_\nu$, but they should also depend on the positions of the other atomic centers $k \neq k_\mu, k_\nu$. In our model, we use an additive scheme

$$\bar{F}_{\mu\nu} = \bar{F}_{\mu\nu,0} + \sum_{k \neq k_\mu, k_\nu} \bar{F}_{\mu\nu,k} \tag{25}$$

where the first leading term is either one- or two-center and the sum runs over the two- or three-center corrections, respectively. The leading one-center integrals are atomic constants

$$\bar{F}_{\mu\nu,0} \overset{k_\mu = k_\nu}{=} \delta_{m_\mu m_\nu} \delta_{l_\mu l_\nu} F_{l_\mu k_\mu}, \tag{26}$$

$$\bar{F}_{\alpha\nu,0} \overset{k_\alpha = k_\nu}{=} 0. \tag{27}$$

The leading two-center integrals

$$\bar{F}_{\mu\nu,0} \overset{k_\mu \neq k_\nu}{=} \sum_m A^{l_\mu}_{m m_\mu}(\mathbf{z}_{\mu\nu}) A^{l_\nu}_{m m_\nu}(\mathbf{z}_{\mu\nu}) F_{m l_\mu l_\nu k_\mu k_\nu}(r_{\mu\nu}) \tag{28}$$

as well as $\bar{F}_{\alpha\nu,0}$ can be reduced to functions of interatomic distances and the well-known transformation matrices $A^l_{mm'}(\mathbf{z})$ of spherical harmonics upon spatial rotation,

$$\mathbf{z}_{\mu\nu} = \mathbf{r}_{\mu\nu}/r_{\mu\nu}, \qquad \mathbf{r}_{\mu\nu} = \mathbf{r}_{k_\mu} - \mathbf{r}_{k_\nu}, \qquad r_{\mu\nu} = |\mathbf{r}_{\mu\nu}|. \tag{29}$$

The two-center corrections to the one-center integrals

$$\bar{F}_{\mu\nu,k} \overset{k_\mu=k_\nu\neq k}{=} \sum_{ml} A^{l_\mu}_{mm_\mu}(\mathbf{z}_{\mu k}) A^{l_\nu}_{mm_\nu}(\mathbf{z}_{\mu k}) B^{l\,l_\mu l_\nu}_{0mm} V_{ll_\mu l_\nu k_\mu k}(r_{\mu k}) \tag{30}$$

as well as $\bar{F}_{\alpha\nu,k}$ are done in the same way, making further use of the triple products $B^{l\ l'\ l''}_{mm'm''}$ of spherical harmonics. The most complicated are the three-center corrections $\bar{F}_{\mu\nu,k}$ with $k_\mu \neq k_\nu \neq k$ that can be exactly reduced only down to nontrivial functions of three variables – we have experimented with the triple series in prolate spheroidal coordinates which can be made very accurate, but in the end we have chosen the (approximate) factorization of the three-center terms

$$\bar{F}_{\mu\nu,k} \overset{k_\mu\neq k_\nu\neq k}{=} \sum_\lambda \delta_{k_\lambda k} S_{\mu\lambda} S_{\nu\lambda} f_\lambda \tag{31}$$

into the products of two-center terms

$$S_{\mu\lambda} \overset{k_\mu\neq k_\lambda}{=} \sum_m A^{l_\mu}_{mm_\mu}(\mathbf{z}_{\mu\lambda}) A^{l_\lambda}_{mm_\lambda}(\mathbf{z}_{\mu\lambda}) S_{ml_\mu l_\lambda k_\mu k_\lambda}(r_{\mu\lambda}) \tag{32}$$

where λ labels the functions of some expansion basis on center k in which the underlying operator is diagonal with eigenvalues f_λ, and $S_{\mu\lambda}$ are the overlap integrals. The three-center terms with one polarization function have minor effect and we set them to zero,

$$\bar{F}_{\alpha\nu,k} \overset{k_\alpha\neq k_\nu\neq k}{=} 0. \tag{33}$$

Of the two-electron integrals, only the one- and two-center ones are of fundamental importance and are included in our model, the rest are generally small and are set to zero. The one-center integrals

$$R^{(0)}_{\mu\nu\mu'\nu'} \overset{k_\mu=k_\nu=k_{\mu'}=k_{\nu'}}{=} R^{(0)}_{\mu\nu\mu'\nu',0} + \sum_{k\neq k_\mu} R^{(0)}_{\mu\nu\mu'\nu',k} \tag{34}$$

have the leading term of atomic constants

$$R^{(0)}_{\mu\nu\mu'\nu',0} \overset{k_\mu=k_\nu=k_{\mu'}=k_{\nu'}}{=} \sum_{ml} B^{l\ l_\mu\ l_\nu}_{mm_\mu m_\nu} B^{l\ l_{\mu'}\ l_{\nu'}}_{mm_{\mu'}m_{\nu'}} G_{ll_\mu l_\nu l_{\mu'}l_{\nu'}k_\mu} \tag{35}$$

and additive corrections for the effect of surrounding atoms

$$R^{(0)}_{\mu\nu\mu'\nu',k} \stackrel{k_\mu=k_\nu=k_{\mu'}=k_{\nu'}\neq k}{=} \delta_{m_\mu m_\nu}\delta_{m_{\mu'} m_{\nu'}}\delta_{l_\mu l_\nu}\delta_{l_{\mu'} l_{\nu'}}V_{l_\mu l_{\mu'} k_\mu k}(r_{\mu k}), \tag{36}$$

for the latter we have also tested the general expression

$$R^{(0)}_{\mu\nu\mu'\nu',k} \stackrel{k_\mu=k_\nu=k_{\mu'}=k_{\nu'}\neq k}{=} \sum_{mm'll'} B^{l\ l_\mu\ l_\nu}_{mm_\mu m_\nu} B^{l'\ l_{\mu'}\ l_{\nu'}}_{m'm_{\mu'}m_{\nu'}} B^{l''l\ l'}_{0\ mm'} V_{l''ll'l_\mu l_\nu l_{\mu'} l_{\nu'} k_\mu k}(r_{\mu k}) \tag{37}$$

but found only the terms with $l'' = l = l' = 0$ to be important, we also find it safe to set

$$R^{(0)}_{\alpha\nu\mu'\nu'} \stackrel{k_\alpha=k_\nu=k_{\mu'}=k_{\nu'}}{=} 0. \tag{38}$$

Of the two-center two-electron integrals, only the long-range ones

$$R^{(0)}_{\mu\nu\mu'\nu'} \stackrel{k_\mu=k_\nu\neq k_{\mu'}=k_{\nu'}}{=} \sum_{\bar{m}\bar{\bar{m}}\bar{m}'\bar{\bar{m}}'mll'} A^{l_\mu}_{\bar{m}m_\mu}(\mathbf{z}_{\mu\mu'})A^{l_\nu}_{\bar{\bar{m}}m_\nu}(\mathbf{z}_{\mu\mu'})A^{l_{\mu'}}_{\bar{m}'m_{\mu'}}(\mathbf{z}_{\mu\mu'})A^{l_{\nu'}}_{\bar{\bar{m}}'m_{\nu'}}(\mathbf{z}_{\mu\mu'}) \times$$

$$\times B^{l\ l_\mu l_\nu}_{m\bar{m}\bar{\bar{m}}} B^{l'\ l_{\mu'}l_{\nu'}}_{m\bar{m}'\bar{\bar{m}}'} G_{mll'l_\mu l_\nu l_{\mu'} l_{\nu'} k_\mu k_{\mu'}}(r_{\mu\mu'}), \tag{39}$$

and in the same way $R^{(0)}_{\alpha\nu\mu'\nu'}$, are included in the model, we studied the effect of other two-center terms and found it safe to set them to zero. We should stress that it is thanks to the special properties of the underlying zero-bond-dipole [28] orthogonalization of the basis functions that all two-electron integrals involving two-center product charge distributions $\phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r})$, $k_\mu \neq k_\nu$, are small and can be neglected. The two-electron dispersion-model integrals are naturally limited to the isotropic two-center terms

$$R^{(6)}_{\mu\nu\mu'\nu'} \stackrel{k_\mu=k_\nu\neq k_{\mu'}=k_{\nu'}}{=} \delta_{\mu\nu}\delta_{\mu'\nu'}G^{(6)}_{l_\mu l_{\mu'} k_\mu k_{\mu'}}(r_{\mu\mu'}). \tag{40}$$

The zeroth-order eigenvalues in Eq. (7) can be optimized as atomic parameters, but we find it enough to set

$$\varepsilon_\mu = \bar{F}_{\mu\mu}, \qquad \varepsilon_\alpha = 1. \tag{41}$$

Now that all molecular integrals in our model are defined in terms of constants and functions of one variable r, parametrized formulas for the latter are needed. The short-range functions of Eqs. (28), (30), (32), and (36) can be fitted to a high accuracy by the general expansion

$$f(r; a, c_0, \ldots, c_n) = \exp(-ar) \sum_{\kappa=0}^{n} (ar)^\kappa c_\kappa \tag{42}$$

11

if enough terms are taken, the long-range functions of Eq. (39) can be done likewise with one long-range term added

$$g_l(r; q, a, c_0, \ldots, c_n) = qr^{-1-l} u_{2l}(ar) + a^{1+l} \exp(-ar) \sum_{\kappa=0}^{n} (ar)^{l+\kappa} c_\kappa \tag{43}$$

with

$$u_n(r) = 1 - \exp(-r) \sum_{m=0}^{n} \frac{r^m}{m!}, \tag{44}$$

and the dispersion tail corrections (40) can easily be added in the form

$$g^{(6)}(r; a, c) = cr^{-6} u_9(ar). \tag{45}$$

We have experimented with these general expansions within our electronic structure model (working with more than 20000 atomic and atom-pair parameters for a set of 7 chemical elements!), but found later that much more compact expressions using sum rules with mostly atomic parameters work quite well and yield more meaningful optimized parameter values. The short-range terms can be given either by a single exponential ($n = 0$) term of Eq. (42) or by a three-parameter hyperbolic-secant function

$$f_o(r; a, b, c) = c \exp(-ar + b) \left(1 + \exp(-2ar + 2b)\right)^{-1} \tag{46}$$

with $b > 0$. For the long-range terms, the leading term of Eq. (43) is enough for all multipole-multipole and charge-multipole interactions and only a single exponential term needs to be added for the charge-charge interactions.

The sum-rule formulas used in this work are as follows. In Eq. (28) we have

$$F_{ml_\mu l_\nu k_\mu k_\nu}(r) = f\left(r; \sqrt{a_\mu a_\nu}\, a_{ml_\mu l_\nu \xi_\mu \xi_\nu}, c_\mu c_\nu c_{ml_\mu l_\nu \xi_\mu \xi_\nu}\right) \tag{47}$$

with the shorthand notation for atomic parameters $a_\mu \equiv a_{l_\mu k_\mu}$, $c_\mu \equiv c_{l_\mu k_\mu}$, and ξ_μ being atom-group labels – in particular, $\xi_\mu = 1$ if μ is on H, $\xi_\mu = 2$ if μ is on a second-row atom Li–F, and $\xi_\mu = 3$ for μ on Na–Cl. Thus we work with atomic parameters and diatomic atom-group parameters $a_{ml_\mu l_\nu \xi_\mu \xi_\nu}$, $c_{ml_\mu l_\nu \xi_\mu \xi_\nu}$ in Eq. (47), and this is also done in the formulas below. For minimal-polarization analog of Eq. (28) we choose

$$F_{ml_\alpha l_\nu k_\alpha k_\nu}(r) = f_o\left(r; \frac{2a_\alpha a_\nu}{a_\alpha + a_\nu} a_{ml_\alpha l_\nu \xi_\alpha \xi_\nu}, b_{\xi_\mu \xi_\nu}, c_\alpha c_\nu c_{ml_\alpha l_\nu \xi_\alpha \xi_\nu}\right) \tag{48}$$

and it should be understood that, here and below, each integral class has its own set of parameters, for example a_ν in Eq. (47) is not the same as a_ν in Eq. (48) – the same notation is used to save space. In Eq. (30) we choose the form

$$V_{ll_\mu l_\nu k_\mu k}(r) = f_\circ\left(r; \frac{2(a_\mu + a_\nu)a_k}{a_\mu + a_\nu + 2a_k}a_{ll_\mu l_\nu \xi_\mu \xi_k}, b_{\xi_\mu \xi_k}, \tfrac{1}{2}(c_\mu + c_\nu)c_k c_{ll_\mu l_\nu \xi_\mu \xi_k}\right) \quad (49)$$

for both minimal-minimal and minimal-polarization integrals. In Eq. (32) we have

$$S_{ml_\mu l_\lambda k_\mu k_\lambda}(r) = f_\circ\left(r; \frac{2a_\mu a_\lambda}{a_\mu + a_\lambda}, b_{\xi_\mu \xi_\lambda}, c_\mu c_\lambda c_{ml_\mu l_\lambda \xi_\mu \xi_\lambda}\right). \quad (50)$$

In Eq. (36) we have

$$V_{l_\mu l_{\mu'} k_\mu k}(r) = f_\circ\left(r; \frac{2(a_\mu + a_{\mu'})a_k}{a_\mu + a_{\mu'} + 2a_k}, b_{\xi_\mu \xi_k}, \tfrac{1}{2}(c_\mu + c_{\mu'})c_k c_{l_\mu l_{\mu'} \xi_\mu \xi_k}\right). \quad (51)$$

In Eq. (39) the form

$$G_{mll'l_\mu l_\nu l_{\mu'} l_{\nu'} k_\mu k_{\mu'}}(r) = g_{l+l'}\left(r; q_{ll_\mu l_\nu} q_{l'l_{\mu'} l_{\nu'}} q_{mll'}, \frac{2(a_\mu + a_\nu)(a_{\mu'} + a_{\nu'})}{a_\mu + a_\nu + a_{\mu'} + a_{\nu'}}, -\tfrac{1}{2}\delta_{0l}\delta_{0l'}\right) \quad (52)$$

is used for both 4-minimal and 1-polarization-3-minimal two-electron integrals, with the atomic multipoles $q_{ll_\mu l_\nu}$ and the fundamental constants of multipole-multipole interaction $q_{mll'}$, in the charge-charge 4-minimal case one exponential term is added. In Eq. (40) we have

$$G^{(6)}_{l_\mu l_{\mu'} k_\mu k_{\mu'}}(r) = g^{(6)}\left(r; \frac{2a_\mu a_{\mu'}}{a_\mu + a_{\mu'}}, c_\mu c_{\mu'}\right). \quad (53)$$

As can be seen, the atomic a-parameters in all these formulas play the role of radial scale factors, the atomic c-parameters are multiplicative prefactors, and the atom-group b-parameters control the shape of functions at shorter distances r.

C. Parametrization procedure

We parametrize our model by minimizing the function

$$\mathcal{E} = \mathcal{E}_E + \mathcal{E}_F + \mathcal{E}_D + \mathcal{E}_g + \mathcal{E}_H + \mathcal{E}_d + \mathcal{E}_a + \mathcal{E}_q + \mathcal{E}_U + \mathcal{E}_6 + \mathcal{E}_c \quad (54)$$

that measures the deviation of molecular properties predicted by the model from the reference values computed by the higher-level theory on a set of molecules (molecular geometries). The energy term

$$\mathcal{E}_E = \sum_n w_n^E \left(\sum_m c_{mn}(\tilde{E}_m - E_m)\right)^2, \quad (55)$$

13

with $m = 1, \ldots, M$ labeling each molecule, can be used not only to fit each predicted energy $\tilde{E}_m$ to the reference value E_m in the trivial case $c_{mn} = \delta_{mn}$, but also for giving higher weights w_n^E to some chemically-meaningful energy differences, such as conformational energy changes, reaction energies and activation barriers.

The effective one-electron Hamiltonian terms

$$\mathcal{E}_F = \sum_m w_m^F \operatorname{tr}\{(\tilde{\mathbf{F}}_m - \mathbf{F}_m)^2\} \tag{56}$$

and the density-matrix terms

$$\mathcal{E}_D = \sum_m w_m^D \operatorname{tr}\{(\tilde{\mathbf{D}}_m - \mathbf{D}_m)^2\} \tag{57}$$

for each molecule take $\tilde{\mathbf{F}}_m$ from Eq. (17) and $\mathbf{F}_m$ from the theory of Section II D below, the density matrices $\tilde{\mathbf{D}}_m$ and $\mathbf{D}_m$ come from the diagonalization of $\tilde{\mathbf{F}}_m$ and $\mathbf{F}_m$. One may put all $w_m^D = 0$ and work with only $w_m^F \neq 0$, assuming that an accurate $\mathbf{F}$-matrix should also yield an accurate $\mathbf{D}$-matrix, but in practice we have not-so-small errors in $\mathbf{F}$-matrices and find it a better compromise to balance between $\mathbf{F}$ and $\mathbf{D}$ terms, thus giving more importance to the occupied-occupied and occupied-virtual blocks of $\mathbf{F}$.

The energy first $\mathbf{g}_m$ and second $\mathbf{H}_m$ derivatives with respect to atomic coordinates for each molecule m are gathered into the terms

$$\mathcal{E}_g = \sum_m w_m^g (\tilde{\mathbf{g}}_m - \mathbf{g}_m)^{\mathrm{T}} \mathbf{W}_m^2 (\tilde{\mathbf{g}}_m - \mathbf{g}_m), \tag{58}$$

$$\mathcal{E}_H = \sum_m w_m^H \operatorname{tr}\left\{\left((\tilde{\mathbf{H}}_m - \mathbf{H}_m)\mathbf{W}_m\right)^2\right\}, \tag{59}$$

where the weight matrix

$$\mathbf{W} = \left(\mathbf{H}^2 + h_0^2 \mathbf{1}\right)^{-1/2} \tag{60}$$

is a well-behaved positive-definite replacement for the inverse $\mathbf{H}^{-1}$. This weighting scheme emphasizes the weak modes and, in our experience, it drives the model towards a balanced reproduction of potential energy surfaces. We set $h_0 = 2^{-10}$ in this work, close to the lowest eigenvalue of $\mathbf{H}$ for the water dimer $(\mathrm{H_2O})_2$ used as a prototype. For reference geometries at stationary points $(\mathbf{g}_m = 0)$, as is normally the case, Eq. (58) can be seen as a weighted sum of squared distances between the stationary points estimated by the model and the reference, whereas Eq. (59) is a weighted sum of squared dimensionless relative deviations of force constant values.

Electric dipole moments and polarizabilities

$$\mathcal{E}_d = \sum_m w_m^d \left| \tilde{\mathbf{d}}_m - \mathbf{d}_m \right|^2, \tag{61}$$

$$\mathcal{E}_a = \sum_m w_m^a \operatorname{tr}\{(\tilde{\mathbf{a}}_m - \mathbf{a}_m)^2\} \tag{62}$$

as well as atomic polar tensors

$$\mathcal{E}_q = \sum_m w_m^q \operatorname{tr}\{(\tilde{\mathbf{q}}_m - \mathbf{q}_m)^{\mathrm{T}}(\tilde{\mathbf{q}}_m - \mathbf{q}_m)\} \tag{63}$$

are included into the optimization in a straightforward way, as are the molecular electrostatic potentials

$$\mathcal{E}_U = \sum_m w_m^U \sum_\kappa \left(\tilde{U}(\mathbf{r}_\kappa) - U(\mathbf{r}_\kappa) \right)^2 s_\kappa \tag{64}$$

on the grids of points $\mathbf{r}_\kappa$ at discretized surfaces with elements s_κ (see also Section II F).

Molecular dispersion coefficients (discussed in Section II E below) are treated in the form of square roots

$$\mathcal{E}_6 = \sum_m w_m^{(6)} \left(\tilde{c}_{6,m}^{1/2} - c_{6,m}^{1/2} \right)^2 \tag{65}$$

as these have a more natural system-size scaling.

The last (optional) term is a sum of soft constraints

$$\mathcal{E}_c = \sum_\kappa w_\kappa^c s \left(\sum_p c_{p\kappa} x_p, \bar{c}_\kappa, \bar{\bar{c}}_\kappa \right), \tag{66}$$

$$s(x, a, b) = \begin{cases} (x-a)^2, & x < a \\ 0, & a < x < b \\ (x-b)^2, & b < x \end{cases} \tag{67}$$

on linear combinations of optimization parameters $\{x_p\}$.

D. Effective Hartree-Fock-like Hamiltonian from correlated wavefunction theory

Given the one-electron nonidempotent density matrices $\mathcal{D}_{\mu\nu\sigma}$ from a correlated wavefunction calculation, with μ, ν labeling (in this Section only) the functions of an extended atomic basis, we set up our Hartree-Fock-like self-consistent field procedure based on the

15

minimization of the energy expression

$$E^{(w)} = \sum_{\mu\nu\sigma} D_{\mu\nu\sigma}^{(w)} H_{\mu\nu} + \frac{1}{2} \sum_{\mu\nu\sigma\mu'\nu'\sigma'} D_{\mu\nu\sigma}^{(w)} R_{\mu\nu\sigma\mu'\nu'\sigma'} D_{\mu'\nu'\sigma'}^{(w)}$$
$$+ \frac{w}{2} \sum_{\mu\nu\mu'\nu'\sigma} \left(D_{\mu\nu\sigma}^{(w)} - \mathcal{D}_{\mu\nu\sigma} \right) R_{\mu\nu\mu'\nu'} \left(D_{\mu'\nu'\sigma}^{(w)} - \mathcal{D}_{\mu'\nu'\sigma} \right) \tag{68}$$

with respect to the idempotent density matrices $D_{\mu\nu\sigma}$. Eq. (68) is a sum of the Hartree-Fock energy and a quadratic density penalty function weighted by w, the associated effective Hamiltonian matrix

$$F_{\mu\nu\sigma}^{(w)} = H_{\mu\nu} + \sum_{\mu'\nu'\sigma'} R_{\mu\nu\sigma\mu'\nu'\sigma'} D_{\mu'\nu'\sigma'}^{(w)} + w \sum_{\mu'\nu'} R_{\mu\nu\mu'\nu'} \left(D_{\mu'\nu'\sigma}^{(w)} - \mathcal{D}_{\mu'\nu'\sigma} \right) \tag{69}$$

is a sum of the Fock matrix and a local spin-dependent correlation potential matrix, the latter arises as the scaled Coulomb potential of the difference spin density. If a complete basis were used, the limit $w \to \infty$, if exists, should yield the exact effective Hamiltonian of Sec. IIB of Kohn and Sham's work [10]. In practice, we work with a finite incomplete basis and a reasonable finite value of w should be chosen. Our procedure can be seen as a nonlocal-exchange-local-correlation analog of the local-exchange-correlation procedure of Zhao, Morrison, and Parr [29].

E. Molecular dispersion coefficients

For two molecules at a large distance r between their centers, the leading asymptotic term of the dispersion part of the intermolecular potential is $c_6 r^{-6}$ with c_6 being a function of the relative orientation. The orientational average of c_6 is a bimolecular constant that can be most easily computed within MP2 as

$$c_6 = \frac{2}{3} \sum_{ai\sigma}^{\text{mol.1}} \sum_{a'i'\sigma'}^{\text{mol.2}} \frac{|\langle \phi_{a\sigma} | \mathbf{r} | \phi_{i\sigma} \rangle|^2 |\langle \phi_{a'\sigma'} | \mathbf{r} | \phi_{i'\sigma'} \rangle|^2}{\epsilon_{i\sigma} - \epsilon_{a\sigma} + \epsilon_{i'\sigma'} - \epsilon_{a'\sigma'}} \tag{70}$$

from the dipole moment integrals over one-electron wavefunctions ϕ and the one-electron energies ϵ, i's label the occupied and a's the virtual states each localized on one of the molecules. Much more complicated expressions can be derived for higher-order correlation methods, but we will use Eq. (70) in this work as it gives accurate enough values for our purpose. We take only the case of two identical molecules and such c_6 becomes a molecular constant of interest, moreover, we have seen that a heteromolecular c_6 is quite close to the

16

geometric mean of two molecular constants. Within our parametrizable model the value to put into Eq. (65) is simply

$$\tilde{c}_6^{1/2} = \sum_{\mu\sigma} D_{\mu\mu\sigma} c_\mu \tag{71}$$

with c_μ as in Eq. (53). By construction, Eq. (53) always gives isotropic c_6 values that follow the rule of geometric mean.

F. Surfaces for sampling molecular electrostatic potentials

A smooth surface is preferable for sampling the molecular electrostatic potential, so we have tried the isodensity surface $\rho(\mathbf{r}) = \rho_0$ and found that $\rho_0 \approx 10^{-6}$ should be used to enclose most of the density as needed. While this may work well for the exact density, the one calculated within a limited finite basis approximation may not be accurate enough at such small values in the tail region, indeed we have seen some rather weird shapes for molecules as simple as LiF.

In this work we set a new definition of molecular surface $p(\mathbf{r}) = p_0$ with

$$p(\mathbf{r}) = \int s(\mathbf{r}' - \mathbf{r})\rho_1(\mathbf{r}', \mathbf{r}'')s(\mathbf{r}'' - \mathbf{r})\, \mathrm{d}^3\mathbf{r}'\, \mathrm{d}^3\mathbf{r}'' \tag{72}$$

in terms of the one-electron density matrix $\rho_1(\mathbf{r}, \mathbf{r})$, the density being its diagonal part $\rho(\mathbf{r}) = \rho_1(\mathbf{r}, \mathbf{r})$, as inspired by the analysis of the exchange repulsion effects. For the broadening function $s(\mathbf{r})$ localized around $\mathbf{r} = 0$ we make the simplest choice

$$s(\mathbf{r}) = c \exp\left(a|\mathbf{r}|^2\right), \tag{73}$$

in the limit $a \to \infty$ with $c = (a/\pi)^{3/2}$ Eqs. (72) and (73) yield $p(\mathbf{r}) \to \rho(\mathbf{r})$. With $c = 1$, $a = 1/16$, and $p_0 = 1/4$ we get good surfaces for the molecules studied in this work. The surface discretization is best done with the spherical quadrature rules [56, 57].

III. CALCULATIONS

The training set of 5581 molecules used in this work covers a broad range of structures built up from 15 elements H, Li–F, Na–Cl, there are both energy-minimum and transition-state geometries of neutral, cationic and anionic species with closed-shell and (high-spin) open-shell electronic configurations dominated by a single determinant. We have spent a lot

of time setting up this database guided by our chemical intuition, finding meaningful structures of (almost) all chemical compositions with up to three nonhydrogen and any number of hydrogen atoms, and less thoroughly for up to eight nonhydrogen atoms. We tried to sample all kinds of chemical bonds – from metallic through covalent to ionic, hydrogen bonded dimers and clusters (for example, neutral, protonated, and deprotonated water clusters with up to 6 O atoms) as well as weaker intermolecular complexes are also carefully chosen.

The reference data were generated by the CCSD method [13] with nonrelativistic Hamiltonian and correlation-consistent atomic basis sets [58] of sizes shown in Table I. The `L1` set has one set of valence polarization functions, the `L11` set is for a correlated treatment of (outermost) core shells, and the `L1+1` set has diffuse functions added as Rydberg shells – this is the entry-level basis for quantitative CCSD calculations, we would have preferred the next-level (`L2`, `L22`, `L2+1`) had we had enough computer power. An archive of all molecular data files is available for download [59].

The model is defined by the energy expression of Section II A, the molecular integrals of Section II B, and the atomic basis set sizes given in Table II with one radial function for each angular symmetry. Besides the full model, 8 simplified models with some classes of integrals set to zero are also analyzed. All parameters of each model have been simultaneously optimized on the full reference data set, the weights assigned to each molecular property and the values of each error term in Eq. (54) at convergence are given in Table III, full details of the settings along with the optimized parameter values can be found in the attached files [59]. Our optimization algorithm computes numerically all first derivatives of molecular properties with respect to the parameters and uses linear searches along the optimal direction to find the nearest local minimum, the choice of the starting guess is so far from straightforward that it cannot be documented here.

It is not trivial to judge the quality of such a model by the net error terms, but still some insight into the role of the whole classes of parametrized molecular integrals can be gotten as they are removed from the model in the order of increasing importance. As expected, the dispersion terms (40) have the smallest impact as shown by Model 2. The role of the polarization basis functions is dramatic – Model 4 that has them all removed nearly doubles the most important error terms, and about 2/3 of this effect come from the one-electron integrals (28) and (30), Model 3 with only the two-electron integrals (39) over the polarization functions included is the simplest one to properly account for the molecular

dipole polarizability but this does also improve the other molecular properties. Starting from Model 4 we can study the role of the integrals over the minimal basis that are new to our model. The two-center corrections to one-center two-electron integrals (36) are the least important, though Model 5 without them shows already about 1.5 greater values of most error functions. Further degradation is caused by zeroing out the multipole moments in the two-center two-electron integrals (52) as in Model 6 and, further on, in the one-center two-electron integrals (35) as in Model 7. At this point, Model 7 is already too simplified to support the burden of the first (58) and second (59) derivative terms as they were before, so we need to relieve it by setting a larger value of h_0 in Model 7′ to get ready for the last strike. A dramatic degradation is seen as the three-center corrections to two-center one-electron integrals (31) are zeroed out in Model 8, and yet again as the two-center corrections to one-center one-electron integrals (30) are dropped in Model 9. With this overview, it should be clear that these integral terms have not only the quantitative but also the qualitative impact on the performance of the model – the simplified models may even fail to reproduce the existence of some stationary point on molecular potential energy surfaces, for example, Model 6 predicts the symmetry of the hydrogen peroxide molecule H_2O_2 to be C_{2h} instead of C_2, and Model 8 gives the cyclobutane molecule C_4H_8 the D_{4h} symmetry instead of the right D_{2d} – the reliable prediction of intra- and intermolecular conformations is a challenge that only our full model seems to meet.

A binary executable code of our program for Linux/x86_64 architecture is made available [59] to the interested researchers worldwide for further testing and some preliminary applications of the new model.

IV.   CONCLUSIONS

We are pleased to have developed the new parametrizable electronic structure model that both has a number of good formal properties and performs well in numerical tests. It evolved through a trial-and-error process until it has reached a satisfactory level of maturity. With this parameter set for 15 elements, many interesting systems can already be studied, we are going to see how it works for the structure prediction of biomolecules and the chemical processes in condensed phase. If some weaknesses will be found, the model can be reparametrized on a more representative training set.

Future work may include: building a comprehensive database of prototype molecular structure for further training and testing of the model, its extension to heavier elements, a reparametrization based on more accurate reference data from CCSD or CCSD(T) calculations with a next-level basis, the investigation of linear-scaling techniques [60–65] for solving the SCF equations of our model to find a both accurate and fast algorithm for studies of large systems. A formalism to treat excited electronic states dominated by single excitations from a single determinant can also be worked out, either within the configuration interaction with single substitutions (CIS) or within the linear response theory.

Acknowledgments

[1] S. Lifson, A. Warshel, J. Chem. Phys. **49**, 5116 (1968).

[2] A. Warshel, M. Levitt, J. Mol. Biol. **103**, 227 (1976).

[3] A. Warshel, R. M. Weiss, J. Am. Chem. Soc. **102**, 6218 (1980).

[4] D. W. Brenner, Phys. Rev. B **42**, 9458 (1990).

[5] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard III, J. Phys. Chem. A **105**, 9396 (2001).

[6] P. Hohenberg, W. Kohn, Phys. Rev. **136**, B864 (1964).

[7] R. Hoffmann, J. Chem. Phys. **39**, 1397 (1963).

[8] X.-P. Li, R. W. Nunes, D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993).

[9] A. H. R. Palser, D. E. Manolopoulos, Phys. Rev. B **58**, 12704 (1998).

[10] W. Kohn, L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[11] C. Møller, M. S. Plesset, Phys. Rev. **46**, 618 (1934).

[12] J. Paldus, J. Čížek, I. Shavitt, Phys. Rev. A **5**, 50 (1972).

[13] G. D. Purvis, R. J. Bartlett, J. Chem. Phys. **76** 1910 (1982).

[14] K. Ragavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, Chem. Phys. Lett. **157**, 479 (1989).

[15] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).

[16] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[17] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[18] H. Iikura, T. Tsuneda, T. Yanai, K. Hirao, J. Chem. Phys. **115**, 3540 (2001).

[19] B. I. Lundqvist, Y. Andersson, H. Shao, S. Chan, D. C. Langreth, Int. J. Quantum Chem. **56**, 247 (1995).

[20] Y. Andersson, D. C. Langreth, B. I. Lundqvist, Phys. Rev. Lett. **76**, 102 (1996).

[21] R. Car, M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).

[22] B. I. Dunlap, J. W. D. Connolly, J. R. Sabin, J. Chem. Phys. **71**, 4993 (1979).

[23] D. N. Laikov, Chem. Phys. Lett. **281**, 151 (1997).

[24] R. A. Friesner, J. Phys. Chem. **92**, 3091 (1988).

[25] R. B. Murphy, Y. Cao, M. D. Beachy, M. N. Ringnalda, R. A. Friesner, J. Chem. Phys. **112**, 10131 (2000).

[26] E. Rudberg, E. H. Rubensson, P. Sałek, J. Chem. Phys. **128**, 184106 (2008).

[27] A. D. Becke, J. Chem. Phys. **107**, 8554 (1997).

[28] D. N. Laikov, Int. J. Quantum Chem. **111**, 2851 (2011).

[29] Q. Zhao, R. C. Morrison, R. G. Parr, Phys. Rev. A **50**, 2138 (1994).

[30] Anonymous Reviewer of the J. Chem. Phys. (2011).

[31] K. F. Freed, Chem. Phys. Lett. **13**, 181 (1972).

[32] S. Iwata, K. F. Freed, J. Chem. Phys. **65**, 1071 (1976).

[33] K. F. Freed, Acc. Chem. Res. **16**, 137 (1983).

[34] C. H. Martin, K. F. Freed, J. Chem. Phys. **105**, 1437 (1996).

[35] J. A. Pople, D. P. Santry, G. A. Segal, J. Chem. Phys. **43**, S129 (1963).

[36] M. J. S. Dewar, Science **187**, 1037 (1975).

[37] M. J. S. Dewar, W. Thiel, J. Am. Chem. Soc. **99**, 4899 (1977).

[38] D. N. Nanda, K. Jug, Theor. Chim. Acta, **57**, 95 (1980).

[39] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, J. Am. Chem. Soc. **107**, 3902 (1985).

[40] J. J. P. Stewart, J. Comp. Chem. **10**, 209 (1989).

[41] M. Y. Filatov, O. V. Gritsenko, G. M. Zhidomirov, Theor. Chim. Acta **72**, 211 (1987).

[42] J. J. P. Stewart, J. Mol. Model. **13**, 1173 (2007).

[43] M. Kolb, W. Thiel, J. Comp. Chem. **14**, 775 (1993).

[44] W. Weber, W. Thiel, Theor. Chem. Acc. **103**, 495 (2000).

[45] J. P. McNamara, I. H. Hillier, Phys. Chem. Chem. Phys. **9**, 2362 (2007).

[46] T. Tuttle, W. Thiel, Phys. Chem. Chem. Phys. **10**, 2159 (2008).

[47] R. Ahlrichs, R. Penco, G. Scoles, Chem. Phys. **19**, 119 (1977).

[48] J. Řezáč, J. Fanfrlík, D. Salahub, P. Hobza, J. Chem. Theory Comput. **5**, 1749 (2009).

[49] M. Korth, J. Chem. Theory Comput. **6**, 3808 (2010).

[50] C. H. Rhee, R. M. Metzger, F. M. Wiygul, J. Chem. Phys. **77**, 899 (1982).

[51] T. J. Giese, D. M. York, J. Chem. Phys. **123**, 164108 (2005).

[52] D. T. Chang, G. K. Schenter, B. C. Garrett, J. Chem. Phys. **128**, 164111 (2008).

[53] W. Liang, M. Head-Gordon, J. Phys. Chem. A **108**, 3206 (2004).

[54] R. P. Steele, M. Head-Gordon, Mol. Phys. **105**, 2455 (2007).

[55] J. Deng, A. T. B. Gilbert, P. M. W. Gill, J. Chem. Phys. **130**, 231101 (2009).

[56] V. I. Lebedev, Zh. Vychisl. Mat. i Mat. Fiz. **16**, 293 (1976).

[57] V. I. Lebedev, D. N. Laikov, Russ. Acad. Sci. Dokl. Math. **59**, 477 (1999).

[58] D. N. Laikov, Chem. Phys. Lett. **416**, 116 (2005).

[59] See EPAPS Document No. [*number will be inserted by publisher*] for reference molecular data, parametrization control settings, optimized parameters, error analysis, as well as the program binary and examples of molecular calculations. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html .

[60] J. J. P. Stewart, Int. J. Quantum Chem. **58**, 133 (1996).

[61] A. D. Daniels, J. M. Millam, G. E. Scuseria, J. Chem. Phys. **107**, 425 (1997).

[62] N. A. Anikin, V. M. Anisimov, V. L. Bugaenko, V. V. Bobrikov, A. M. Andreyev, J. Chem. Phys. **121**, 1266 (2004).

[63] T.-S. Lee, D. M. York, W. Yang, J. Chem. Phys. **105**, 2744 (1996).

[64] S. L. Dixon, K. M. Merz, Jr., J. Chem. Phys. **104**, 6643 (1996).

[65] A. D. Daniels, G. E. Scuseria, J. Chem. Phys. **110**, 1321 (1999).

TABLE I: Atomic basis set sizes for CCSD calculations.

Atom	Set	Core	Contracted	Primitive
H	L1		2,1	8,4
Li, Be	L11		4,3,1	12,8,4
B, C	L1	1	3,2,1	12,8,4
N, O, F	L1+1	1	4,3,1	14,10,4
Na, Mg	L11	1	5,4,2	18,13,8
Al, Si, P	L1	2,1	4,3,1	18,13,5
S, Cl	L1+1	2,1	5,4,1	20,15,5

Number of radial functions for each angular quantum number is given.

TABLE II: Atomic basis set sizes in the parametrized model.

Atom	l_μ^{max}	l_α^{max}	l_λ^{max}
H	0	1	1
Li, Be	1	1	2
B, C, N, O, F	1	2	2
Na, Mg	1	1	2
Al, Si, P, S, Cl	1	2	2

Maximum angular quantum number in the minimal l_μ^{max}, polarization l_α^{max}, and projection (31) l_λ^{max} basis.

TABLE III: Errors in molecular properties from the parametrized models.

Integral term[a]	Eq.	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 7′	Model 8	Model 9
$\bar{F}_{\mu\nu,k}^{(k_\mu=k_\nu\neq k)}$	(30)	+	+	+	+	+	+	+	+	+	
$\bar{F}_{\mu\nu,k}^{(k_\mu\neq k_\nu\neq k)}$	(31)	+	+	+	+	+	+	+	+		
$G_{ll_\mu l_\nu l_{\mu'} l_{\nu'},k_\mu}^{(l>0,k_\mu=k_\nu=k_{\mu'}=k_{\nu'})}$	(35)	+	+	+	+	+	+				
$q_{ll_\mu l_\nu}^{(l>0)}$	(52)	+	+	+	+	+					
$R_{\mu\nu\mu'\nu',k}^{(0)(k_\mu=k_\nu=k_{\mu'}=k_{\nu'}\neq k)}$	(36)	+	+	+	+						
$\bar{F}_{\alpha\nu,0}^{(k_\alpha\neq k_\nu)}$	(28)	+	+								
$\bar{F}_{\alpha\nu,k}^{(k_\alpha=k_\nu\neq k)}$	(30)	+	+								
$R_{\alpha\nu\mu'\nu',0}^{(0)(k_\alpha=k_\nu\neq k_{\mu'}=k_{\nu'})}$	(39)	+	+	+							
$R_{\mu\nu\mu'\nu'}^{(6)}$	(40)	+		+	+						
Number of parameters[b]		720	690	524	513	434	406	378	378	264	185

Error term[c]	Eq.	w [d]										
$\mathcal{E}_{E_{\text{total}}}$	(55)	2^{13}	975.8	1013.0	1718.5	2206.1	3307.7	5006.1	5940.6	4498.5	8176.7	12026.6
$\mathcal{E}_{E_{\text{react}}}$	(55)	2^{22}	228.1	255.1	497.8	562.6	1022.6	1459.8	1662.8	591.4	3010.9	4251.5
$\mathcal{E}_F$	(56)	2^{-2}	998.3	989.6	1131.8	1190.0	1013.9	1107.5	1266.4	902.5	1684.5	1811.7
$\mathcal{E}_D$	(57)	1	267.2	283.1	522.7	552.3	675.7	771.5	1042.8	720.9	1318.9	1115.3
$\mathcal{E}_g$	(58)	2^5	3194.3	3255.2	5067.6	5978.6	9712.6	12905.1	13887.4	3432.9	5439.0	8111.3
$\mathcal{E}_H$	(59)	1	1961.9	1987.2	3398.4	3957.8	6313.1	7774.1	9329.4	1872.7	2990.5	4356.6
$\mathcal{E}_d$	(61)	2^{-3}	7.5	7.6	11.7	22.6	31.5	56.9	63.6	47.2	86.7	119.7
$\mathcal{E}_a$	(62)	2^{-10}	62.6	58.0	70.0	–	–	–	–	–	–	–
$\mathcal{E}_q$	(63)	1	196.1	199.9	289.3	477.6	615.7	661.4	782.9	669.1	1315.5	1258.3
$\mathcal{E}_U$	(64)	2^3	324.3	324.5	332.2	331.6	305.9	262.3	255.6	278.5	263.5	232.2
$\mathcal{E}_6$	(65)	2^{-3}	85.4	–	200.4	284.2	–	–	–	–	–	–
	(60)	h_0 [e]	2^{-10}	2^{-10}	2^{-10}	2^{-10}	2^{-10}	2^{-10}	2^{-10}	2^{-5}	2^{-5}	2^{-5}

[a] The integral terms of the Eqs. given are included in the model if marked by the "+" sign, otherwise set to zero.

[b] The number of independent parameters optimized for each model.

[c] The error terms of the Eqs. given with [d] the assigned weights and [e] the h_0 values of Eq. (60) are listed (in au) for each model if included in the optimization, otherwise assigned zero weight and marked by "–".

JCP Manuscript **A11.06.0280**

Title: **A new parametrizable model of molecular electronic structure**

Author: **Dimitri N. Laikov**

Discussion of Reviewers' Comments and Changes made to the Maniscript

Many thanks to the Reviewers for the careful reading of the Manuscript and the valuable comments and suggestions. I have an impression that my work has excited a lot of interest – the Reviewers seem to be eager to see more details on the performance of the new model, so they both ask for more tests of the model's accuracy to be done and reported right now. I strongly believe that such extended tests would make up a very good report publishable separately, not to say that I have limited time left would I have wished to do it now, moreover, I would be delighted to see such testing done by a third party. I find a good way to resolve this problem – the Supplementary Information will now include the binary executable file (single-threaded only) of my computer code for the most common machine architecture (Linux/x86_64) that can be used by anyone in the world to run molecular calculations with the new model – geometry optimizations of stable and transition states, harmonic vibrational analyses, intrinsic reaction coordinates – sample input files are also included and an input description is added to the README file. On the other hand, the Reviewers found no flaw in the most fundamental part of the work – the mathematics of the new model set forth in the 73 Equations of the Manuscript. The now highly cited 1963 paper of Pople, Santry, and Segal (Ref. 35) comes immediately to my mind – it included only the mathematical equations and *no numerical tests at all*. I do include the Calculations section to show that the new model is indeed *parametrizable* and that the new terms added to the energy expression of the model lead to great improvements in the computed molecular properties.

I have revised the Manuscript to improve its quality. The Reviewers' comments follow below (in typewriter font) with my answers and an account of the changes made.

```
Reviewer #1 Evaluations:
RECOMMENDATION: Publish in JCP with mandatory revision (minor)


See attached file.
```

Referee report on A11.06.0280

Title: A new parametrizable model of molecular electronic structure

Author: Dimitri Laikov

In this paper, a heavily parametrized semi-empirical model
is presented with significantly better accuracy than classic semi-
empirical approaches.
I find this topic should be interesting to the JCP readers and when
properly presented would deserve a publication in JCP.
However, the paper should be significantly improved before acceptance.

1. My major issue is the quality of the Section III. The section lacks
depth. This section is most important for users of the model, yet
a lot of details are not explained.

As I noted before, the Calculations (Section III) are of less importance than the Theory
(Sections II) in this work, but I have made some changes and additions to address this issue
(see also below).

The only major data of the paper in Table III are poorly presented.
It took me some time to understand that these are weighted errors
multiplied by "w" factor but only after digging into the details
of the Supplementary Info knowing some of the results I guessed
the energy units are hartree. The table should be described in depth,
the units explained and all the symbols either explained in the caption
or related to the relevant section in the text.

Footnotes are now added to Table III to explain the data, the weighted errors $\mathcal{E}_E$, $\mathcal{E}_F$, $\mathcal{E}_D$,
$\mathcal{E}_g$, $\mathcal{E}_H$, $\mathcal{E}_d$, $\mathcal{E}_a$, $\mathcal{E}_q$, $\mathcal{E}_U$, $\mathcal{E}_6$ in Table III were meant to be those defined in the text, but to

make it clearer one more column is now added to the table with equation numbers for each of these terms, the use of atomic units is stated.

```
2. Nevertheless the raw errors for the training set are still not very
useful for
a general reader. The data should be related to some other methods
and the simplest way to achieve it is to utilize some precomputed
databases. I recommend the GMTKN30 database [J. Chem. Theory Comput., 7,
291-309 (2011)]
with the data available on this webpage:
\protect\vrule width0pt\protect\href{http://toc.uni-muenster.de/GMTKN/GMTKN30/GMTKN30mai
The database is comprehensive, well organized and the
geometry information is machine readable, however,
other databases can also be used. The complete
database would be the most useful but a relevant subset covering
a wide range of chemical problems should also be sufficient. Mean errors,
mean absolute errors and maximal errors would show bias, average accuracy
and
worst case accuracy of the method and could be compared to some DFT
or wavefunction-based methods. Also, it would show how well the method
is transferable to systems not used in the training set. The errors
in the reaction energies from the Supporting Information are very good
indeed,
usually below 1 kcal/mol. Of course, some deterioration of the accuracy
is
inevitable, partly due to rather limited accuracy of the training set,
however it would show how useful is the method at this stage.
```

Thanks to the Reviewer for drawing our attention to the GMTKN30 database built up by the outstanding researchers from the Organic Chemistry Institute of the Westfälische Wilhelms-Universität Münster, Germany. I have visualized all 1218 molecular geometries from the GMTKN30 database – *ihre Bilder habe ich vor mein geistiges Auge gestellt* – and my trained eye tells me that there is likely less diversity in that set than in my own training

set of 5581 molecules! Yes, there are some bigger molecules among those 1218, but they are built up from some rather common fragments. As I noted before, the Calculations (Section III) of my work were intended mainly to show the *parametrizability* of the new model and the more extended tests should be done and reported *separately*, preferably by a third party. The first-generation parametrization of the model reported in this work is based on the reference data from the CCSD method with an entry-level basis set, so I would expect the atomization energies (and even the bond-breaking reaction energies) to agree rather poorly with any complete-basis-set extrapolated CCSD(T) or experimental thermochemical data. Other molecular properties – first of all the geometries, but also the concerted reaction energies, proton affinities, and hopefully even the harmonic vibrational frequencies may be already quite meaningful.

I decided to do much more than just the benchmarking of the model on some well-established molecular database – I provide a binary executable code of my program (in the Supplementary Information) that can be used by anyone in the world to do molecular calculations with the new model such as (but not limited to) the testing and benchmarking of the model's accuracy, and I am looking forward to get any useful feedback, either privately or in a published form.

```
3. The README in the Supporting Information should also be extended. Some
of the information in the files is self-explanatory
(e.g. errors) but I find others (e.g. some info from the files in the mol
directory)
very cryptic.
```

The README file is now greatly extended to give detailed information on the format of the data files.

```
4. The flow of the description in Section II is chaotic. Some
of the symbols are defined long after they are first used which
greatly impairs the reading. For instance, after Eq. (3),
there is a discussion of E_2 term before all the symbols in
Eq. (3) are explained. Moreover, section IIC uses some
symbols that are only defined in sections IID and IIE.
```

I find no better way to write the equations of the Theory (Section II). May the Readers forgive me this imperfection, although I find it quite natural to explain some terms later, this seems to me to be the best compromise.

5. How were the parameters optimized? With non-linear parameters, there are
several choices of methods. How the optimization was guided not to get
stuck in a local minimum?

I have added one sentence about the optimization algorithm to Section III. Needless to say that for such a heavy optimization problem I have no means to make global searches, I can only hope to be lucky enough to find the global minimum by intuition and painful trial-and-error work, or at least to find a good local minimum!

6. It would help the readers if the "lengthy but straightforward"
expressions on pages 7 and 8 were put in the main text or the
Supporting Information.

I think that even an average undergraduate student in physical sciences should be able to derive these expressions – the derivatives of the analytical formulas with only additions, subtractions, multiplications, and divisions, however lengthy they may be.

7. Is the computer code described on page 8 going to be distributed?
If the author plans to distribute the code, the reference in the paper
would help the readers finding the code.

I put a single-threaded binary executable code into the Supplementary Information, along with an input description and sample input and output files. Those who are interested in a deeper work with the code may try to contact me personally for a collaboration.

Reviewer #2 Evaluations:
RECOMMENDATION: Reconsider for JCP after mandatory revision (major)

Reviewer #2 (Comments to the Author):

The manuscripts reports an impressive amount of work that is aimed
at the development of a new general-purpose parametrizable model
for electronic structure calculations. Key ideas are the use
of a minimal basis set (MBS) of orthogonalized atom-centered basis
functions for the SCF treatment, the inclusion of polarization effects
in the spirit of second-order perturbation theory, parametrized expressions
for the one-electron and two-electron integrals (with the neglect
of many small terms), an elaborate parametrization procedure
that covers many different properties and employs CCSD reference data,
the extraction of an effective MBS Hamiltonian from CCSD calculations,
and the inclusion of dispersion effects into the two-electron integrals
via an MP2-related procedure. The parametrization is carried out
for a diverse and comprehensive basis set of 5581 molecules,
and overall error measures for the molecular properties are reported
for the recommended full model and 9 simplified variants.
The EPAPS material contains more than 879,000 lines of computer-generated
data.

In the Introduction, the author gives a nice general overview over the field.
In spite of the need to be selective, he should cite and briefly discuss
the work of K. F. Freed who has written more than 20 papers between 1972
and 1996 on the derivation of semiempirical MBS methods (in particular
of the MBS integrals) from rigorous ab initio theory, in a spirit that is
somewhat similar to the present work; leading references could be
Acc. Chem. Res. 1983, 16, 137 (an early review) and JCP 1996, 105, 1437
(one of the latest papers).

Thanks to the Reviewer for making me aware of Prof. Karl F. Freed's work, now I have
read his key publications and I have added 4 citations of his work (Ref. 31-34). I believe it
is a good idea to cite that work, although those model Hamiltonians are of a different kind
– to be used in a fully correlated theory (full CI or an approximation thereof).

The statement on page 4 that the orthogonalization-corrected methods

from ref 38-39 "did not lead to much improvement" over previous
semiempirical methods also needs to be corrected, for detailed evidence
see the following publications: JPCA 2007, 111, 5751; JCTC 2010, 6, 1546;
JCTC 2011, 7, online (dx.doi.org/10.1021/ct200434a).

I have now corrected that statement to "lead only to a limited improvement" which I believe
should be more diplomatic.

I have read the abstracts of the suggested publications:

1. *Looking at Self-Consistent-Charge Density Functional Tight Binding from a Semiempirical Perspective*, Nikolaj Otte, Mirjam Scholten, and Walter Thiel, J. Phys. Chem. A, 2007, 111 (26), pp 5751–5755.

2. *Benchmark of Electronically Excited States for Semiempirical Methods: MNDO, AM1, PM3, OM1, OM2, OM3, INDO/S, and INDO/S2*, Mario R. Silva-Junior and Walter Thiel, J. Chem. Theory Comput., 2010, 6 (5), pp 1546-1564.

3. *Benchmarking Semiempirical Methods for Thermochemistry, Kinetics, and Noncovalent Interactions: OMx Methods Are Almost As Accurate and Robust As DFT-GGA Methods for Organic Molecules*, Martin Korth and Walter Thiel, J. Chem. Theory Comput., Article ASAP.

I find Paper 1 to be of little relevance to my work as it deals, first of all, with a different type
of models (SCC-DFTB) that I am not willing to criticize. Paper 2 deals with electronically-
excited states and is thus also from a different field. Paper 3 may be indeed interesting,
although it deals only with molecules built up from 4 elements H, C, N, and O, moreover, I
have no full-text access to JCTC at my institution now, asking a friend in the US to send
me a copy of that paper would have taken some time because he is in vacation. Thus I
cannot cite any of the 3 Papers.

The Theory section contains many interesting and novel ideas.
The proposed formalism clearly goes beyond the currently available
semiempirical schemes and accounts for a number of further interactions.
Having said this, one should also point out that this formalism still
employs many approximations and neglects many terms (which is mandatory

when aiming for a very fast electronic structure method). For example,
the method still uses a minimal basis, polarization is treated
at second order with some empirical damping, numerous approximations
are adopted for the molecular integrals (see II.B; use of additive schemes
and factorizations and setting small terms to zero and fitting
the remaining terms, etc), there is a fairly extensive parametrization,
and the reference ab initio methods are CCSD with a rather modest basis
(in general) and MP2 (for modeling the leading dispersion terms).
I find many of the methodological ideas interesting, but from the description
in the manuscript, it is very difficult for me to judge the soundness and
quality of the approximations in the proposed scheme. Generally speaking,
the reader would like to see more detailed justifications on the choices
and approximations made; however, I do not see how this can be done easily.
One possibility might be to give some numerical examples and integral plots
for prototypical interactions and small molecules, but I admit that this
could easily grow into a separate paper.

I have though a lot in the last years on the foundation and justification of parametrizable electronic structure models and made many numerical studies on the behavior of the underlying molecular integrals that can be computed in some way from first principles. I decided not to go into any such details in this work as it would have easily overwhelmed everything else to a point that hardly anyone in the world would have ever been able to read the whole text! The Reviewer admits rightly that "this could easily grow into a separate paper".

The Calculations section summarizes massive computational work
on the parametrization of 15 elements: H, Li-F, and Na-Cl.
The accuracy of the results will be limited by the use
of CCSD/TZVP-type reference data (on top of the other approximations,
see above). The quality of the results is addressed only very summarily
by giving overall errors for various properties (Table III). This is done
in a way that baffles me. Even if I assume that the data in Table III
are given in atomic units (not specified), I do not really understand
what the numbers mean. I do not see answers to the following simple

```
questions: What is the typical error of the best (full) model
for energies, geometries, dipole moments, polarizabilities, etc?
Are the errors fairly uniform for different elements or not?
How does the method perform for standard benchmark sets such as G2 and G3?
```

I have improved Table III according to the comments of Reviewer #1, the use of atomic units is now noted. The main point of Table III is to show how the accuracy of the model changes as the new terms are added or removed. The Reader can see the typical errors for energies in the listings (file "energy.dat") found in the Supplemetary Information, the weighted errors for all molecular properties studied for each molecule are also listed (file "errors.dat"), these data can be easily imported by the interested Reader into a spreadsheet and analyzed in many ways! Because this first-generation parametrization of the new model is based on the reference data from CCSD calculation with an entry-level basis set, the atomization energies would agree quite poorly with the experimental data, but the more chemically-interesting reaction energies should already be meaningful. I find the reproduction of molecular geometries, especially for weakly-bound intermolecular complexes, to be a greater challenge than even the accurate total energies for typical molecules, so my parametrization functional of Section IIC is constructed to give these properties a greater boost, as discussed in the text.

To let anyone see how the model works for any molecular system of interest, I have now included the binary executable code of my program into the Supplementary Information, as noted above in the answer to the comments of Reviewer #1.

```
The EPAPS material is also essentially inaccessible to me. I appreciate
that the almost 880,000 lines of computer-generated data might contain
all the answers to my questions, but I am afraid that nobody other than
the author will be willing to check and digest this material.
```

I have now added a detailed description of the file formats so that these data can be easily analyzed – if not by hand then using some simple parser program that even an average undergraduate student in physical sciences should be able to write if needed. To visualize all 5581 molecular structure of the training set one can run, for example, this command (under UNIX bash shell in an X-Window terminal) in the `data/` directory:

```
for m in mol/*.in ; do bin/xm $m ; done
```

The binary of `xm` viewer is provided in the Supplementary Information.

```
What to do with this paper? On the one hand, I see potentially interesting
ideas, but on the other hand, my dominant feeling is that I cannot really
judge the quality of the approximations made nor the quality of the results.
In my opinion, it would be a disservice to the author to publish
this potentially important manuscript as is.
```

I find this statement too pessimistic. Ideally, to "really judge the quality of the approximations made" and "the quality of the results", an independent reproduction of all the results of my work would be needed, but this cannot be done quickly (in a few months) even by the bravest worker in this field. I fear a much greater "disservice to the author" would have been if the Manuscript were rejected on the maximalist grounds.

```
I suggest major revision. Apart from the explicit points mentioned above,
the author should be required to improve the manuscript with regard
to the following:
(a) better justification of the approximations made;
```

This point is already discussed (see above).

```
(b) better documentation of the quality of the results, including
a statistical evaluation at least with regard to the G2/G3 and S22
benchmark sets, and preferably also with regard to other established
benchmarks;
```

My training set of 5581 molecules is so diverse that these benchmark sets would hardly show anything new. Most molecules of the G2 set are already found in my set and it is nearly so for many other benchmarks. The traditional semiempirical models have rather few parameters and used to be parametrized on a small training set and then an extensive testing was very much needed to see their reliability. The situation is very different in my case where I aim at the highest imaginable diversity already in the training set.

```
(c) some brief information on practical matters, e.g., cpu times relative
to standard semiempirical and DFT methods;
```

The CPU times would depend on the implementation, but this model should be only marginally slower (less than twice) than the traditional semiempirical models, and this can be seen simply from the equations of the model, where the bottleneck is nearly the same linear algebra with the matrices of the same minimal-basis dimension.

```
(d) adding a detailed table of content and detailed explanation
to the EPAPS material so that readers can appreciate the documented results.
```

This is done as noted above.

I hope to have made the important changes to improve the quality of the Manuscript. Many thanks to everyone for the attention.